DOCUMENT RESUME

ED 435 753 TM 030 405

AUTHOR Le Rouzie, Violaine; Ouchi, Fumika; Zhou, Chunnong
TITLE Measuring "What People Learned" versus "What People Say They Learned": Does the Difference Matter?
PUB DATE 1999-11-00
NOTE 10p.; Paper presented at the Annual Meeting of the American Evaluation Association (Orlando, FL, November 3-6, 1999). Colored figures may not reproduce well.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adults; Correlation; Employees; *Learning; *Pretests Posttests; *Self Evaluation (Individuals); *Training
IDENTIFIERS Self Report Measures; *World Bank

ABSTRACT

Training courses organized by the World Bank Institute (WBI) have recently started to assess participant learning using a randomized, cognitive pretest-posttest. Some trainers, however, feel reluctant to use this Level 2 evaluation (D. Kirkpatrick, 1994) in their courses, and continue to rely on participants' self-assessment of their own knowledge, a more common, traditional approach, that has been used by the WBI. This study investigated whether participants' perceptions about what they have learned can be a valid proxy of what they have actually learned. In some cases self-assessment by participants was positively correlated with the amount of their actual learning. However, the correlation was too weak to enable researchers to rely on perceived self-reported data on learning to measure actual learning. Breaking down the data by gender, region, education, or years of related experience showed that none of the groups studied were consistently able to validly assess how much they learned in a course. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 435 753

Measuring 'What People Learned' versus 'What People Say They Learned': Does the difference Matter?

Violaine Le Rouzie
Fumika Ouchi
Chunnong Zhou

BEST COPY AVAILABLE

TM030405

Presented at the
American Evaluation Association Annual Conference 1999
Chicago, November 1999

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

V. Le Rouzie

Measuring 'What People Learned' versus 'What People Say They Learned': Does the Difference Matter?

Authors: Violaine Le Rouzic, Fumika Ouchi, and Chunnong Zhou
Evaluation Unit of the World Bank Institute

Presented at the
American Evaluation Association Annual Conference 1999
in Orlando on November 5, 1999

Abstract

Training courses organized by the World Bank Institute (WBI) have recently started to assess participant learning using a randomized, cognitive pre-post test[1]. Some trainers, however, feel reluctant to use this Level 2[2] evaluation in their courses, and continue to rely on participants' self-assessment of own knowledge—a more common traditional approach that has been used in WBI. This study investigated whether participants' perceptions about what they have learned can be a valid proxy of what they have actually learned. We found that in some cases self-assessment by participants (i.e., trainees) was positively correlated with the amount of their actual learning. However, the correlation was too weak to enable us to rely on perceived self-reported data on learning to measure actual learning. Breaking down the data by gender, region, education or years of related experience showed that none of the groups studied were consistently able to validly assess how much they learned in a course.

Introduction

Traditionally, all training courses offered by WBI have been using some types of a course evaluation by participants to assess their performance. Until a few years ago, the most common form of course evaluation was simply an end-of-course evaluation form that asked participants for their reaction to the course. Then the idea of measuring participant learning was introduced, as part of the effort to assess the effectiveness of the Institute's training courses on participants.

Because of the initial resistance of most trainers to directly assess participants' learning, the Institute first used a 'post-then self-assessment' evaluation method. This method consists of asking participants to assess their level of knowledge of the course topics, based on their perceptions. For this, participants complete a single questionnaire at the end of the course. They give two ratings for each topic of the course. With the first rating, participants **retrospectively** assess of how much they knew the topics just before the course. The second rating asks participants to self-assess their **current** knowledge of the course topics.

---

[1] In this paper, the words 'test' and 'assessment of actual knowledge' are used interchangeably.

[2] In this paper, evaluation levels refer to the four Levels of Kirkpatrick. Level 1 measures participants/trainees' reactions to the training, and Level 2 their actual learning of the training materials. In our context, both measures are collected via questionnaires administered to the full group of participants at the training course.

This self-assessment method helped to make the idea of measuring learning more acceptable among WBI's trainers. It eventually enabled the Evaluation Unit to implement a method that would objectively measure the levels of knowledge acquisition by participants. A 'randomized, cognitive pre-post assessment of actual learning' method was designed and introduced in 1998.

In this Level 2 evaluation method, trainers develop a set of multiple-choice questions (usually 30). The Evaluation Unit randomly assigns them to two groups of questions. Half of the questions is answered by the participants at the beginning of the course, and the other half at the end of the course. The assessments are anonymous. Confidential numbers are used to match individual responses to the pre-course and post-course assessment. Participants are instructed to choose one response option per question. If they don't know the answer, they should simply answer 'I don't know.' This procedure is not a test of the participants' knowledge, as it would be given in a university, but an assessment of the trainers' ability to effectively convey the course contents to the participants.

However, some trainers have expressed a reluctance to move to the Level 2 evaluation and have remained with a Level 1 evaluation, relying on participant 'post-then self-assessment' of knowledge. Their alleged major reason for being reluctant is that many courses target senior-level government officials from various countries, and the trainers simply don't feel that it is appropriate to give a 'test' to this kind of senior-level audience at their courses.

At the Evaluation Unit of the Institute, the reluctance of some trainers to move to the Level 2 evaluation has been an issue. For those courses that used only the Level 1 evaluation with a post-then method, we have been forced to base our judgement of participant learning solely on self-assessed knowledge gains reported by course participants. There are no data available for these courses that would help us objectively assess potential changes in participant knowledge. Despite this lack of data, some trainers started to report perceived learning gains obtained through self-assessment methods as if they were actual learning gains. For these reasons, we felt that it was important to find out whether or not self-reported data from participants can serve as a 'proxy' for measuring learning.

We looked into the literature and found various studies that investigated the potential relationship between training attributes and/or outcomes, such as changes in work behaviors or knowledge test scores, and trainee characteristics or their reactions, e.g., Faerman and Ban (1992), Warr and Bunce (1995), Alliger et al (1997). While the studies all appear to conclude that trainee *reactions*, in particular, cannot be used as a substitute for actual learning, some of them did report examples where positive correlations existed between trainee reactions and learning outcomes in their studies (e.g. Alliger et al).

Alliger et al (1997) slightly modified the Kirkpatrick's four levels of evaluation, and provided a more detailed breakdown of training criteria. They discussed a case where utility-type reaction measures (reactions to such questions as 'To what degree will this training influence your ability later to perform your job?,' 'Was this training job

relevant?,' and 'Was the training of practical value?') did correlate with immediate learning and on-the-job performance. They also found that this correlation was much stronger than that observed with affective-type reaction measures (e.g. 'liking of training'). They reported that trainee reactions should not be used blindly for the assessment of actual learning, but utility questions can be used as the better estimate if the training criterion is carefully defined and developed.

Nancy Dixon's study in 1990 may be the closest study to ours. She asked the question: 'To what extent do perceptions of amount learned correlate with actual learning scores?' Her results using post-test results of 1,200 employees of a large manufacturing company in the Southwest (US) was that there was no relationship between participants' perceptions of how much they learned and their actual test scores.

The literature being rather slim and inconclusive so far, we decided to use our data set (collected from training courses offered by WBI) to find out what it would tell us about the relationship between participant reactions and their actual learning. We approached this study from different perspectives, using three different methods:

1) Comparison of participant ratings to the traditional end-of-course Level 1 question: 'Extent to which you have acquired information that is *new* to you' with their scores on pre-post actual knowledge assessments;
2) Comparison between learning gains reported by participants based on post-then self-assessment and learning gains observed on pre-post assessment of actual learning; and
3) Observation of the participants' responses to the pre-post actual knowledge assessment from the point of view of their ability to perceive whether or not they knew the answer to a question. For this, we compared the participants who correctly assessed their knowledge of the question by selecting either the right answer or by selecting 'I don't know' with the participants who selected a wrong answer to the question.

Evaluation Methods and Respective Preliminary Results

First we need to emphasize the fact that our data set is small and that our preliminary findings need to be strengthened by additional data. Since we build our data according to what trainers want to find out on each specific course, no question (be it demographic, Level 1 or Level 2) is asked systematically. Consequently, breaking down our data by demographic categories most often results in very small samples. This is important to keep in mind while reading the results provided throughout this study.

The study tried to answer the key question: 'Can measuring people's perceptions on learning be a valid proxy for measuring what they have learned?' from the following three different methods.

Method I compared participant ratings to the traditional end-of-course Level 1 question: 'Extent to which you have acquired information that is *new* to you' with their pre-post test scores. The data included the responses of 189 participants in six courses.

To find a relationship between how much participants actually learned during the courses, and how much they felt they learned, we observed the Pearson product-moment correlation between the learning gains and the new information acquired.

The 'learning gain' was measured by the difference between the percentage of correct answers in the post-course actual knowledge assessment and the percentage of correct answers in the pre-course actual knowledge assessment.

The answer to the question 'extent to which you have acquired information that is new to you' was used as a proxy for what participants felt they learned. This question was rated on a scale of 1 to 5 where 1 means 'minimum' and 5 means 'maximum.'

The result showed a significantly positive relationship between these two measures ($p = 0.039$). However, the relationship was weak: $r = 0.150$ ($N = 189$).[3]

Then we observed the relationship among sub-groups of our data set to see if some demographic groups were more likely than others to perceive the extent to which they learned the contents of the course. The demographic variables observed are formal educational level[4], years of related work experience, region of origin[5], and gender. We examined the results of each individual subgroup (e.g., male) as well as the combination of sub-groups (e.g., male from South Asia).

We found that participants who had completed at least a Masters Degree and had less than five years of work experience in the field of the course were more likely to self-assess accurately how much they learned during the course. The correlation coefficient was 0.588 with a significance level of 0.003. However, only 24 participants met both these criteria in our data set.

With this only exception, none of the other demographic groups observed showed a significant relationship between what they learned and what they said they learned.

---

[3] Throughout this study, "p" represents the two-tailed significance level; "r" the Pearson product-moment correlation coefficient, and "N" the number of participants in the computation.

[4] Our educational data was composed of four groups: 1) people with less than a Bachelor degree, 2) with a Bachelors degree or equivalent, 3) with a Masters degree or equivalent, and 4) with a Doctorate or higher.

[5] The regions referred to the wide geographic divisions used by the World Bank: Sub-Saharan Africa, East Asia and Pacific (except, Australia, Japan, New Zealand and South Korea); Eastern Europe and Central Asia, Latin America and the Caribbean, Middle East and North Africa, and South Asia. Grouped under 'other regions' are Australia, Japan, New Zealand, South Korea, North America and Western Europe.

Method II compared learning gains reported by participants based on post-then self-assessment with learning gains observed on pre-post assessment of actual learning.
Only one of our courses used both participant post-then self-assessment of knowledge and cognitive pre-post assessment of knowledge. Forty-eight participants attended the course. Our evaluations are anonymous, yet we need to be able to link participants' responses to questionnaires administered at different points in time. Therefore we use numbers to match participants' pre-course responses with their post-course responses on an individual's basis. In the course observed, thirty-nine participants were matched and therefore used for this study.

First, we compared the average self-assessed learning gain ratings with the average actual learning gain scores. To simplify the comparison, we converted our scales to make them comparable. The 15 questions asked in the self-assessment part of the evaluation questionnaires used a scale of '1' to '5' where '1' meant 'very low' and '5' meant 'very high.' The actual knowledge assessment asked 15 questions before the course and 15 questions after the course. Each question (if answered correctly) earned one point, making the possible scores to the pre-post assessment of actual learning range from 0 to 15[6].

Results: Out of the scale of 100. (N=39)

	Self-assessed	Actual
Pre	49.4	46.0
Post	73.6	44.1
Learning (post-pre)	24.2	-1.9

Second, we correlated the average pre-course self-assessed level of knowledge with the pre-course actual assessment score pooling together all questions and taking each participant as a unit. We then repeated the same operation with the post-course data. Finally, we correlated perceived learning ratings with actual learning scores in the same fashion. (The learning ratings were obtained by deducting the average pre-course ratings from the average post-course ratings. Learning scores were computed following the same principles.)

In two of the three computations, the results were significant:

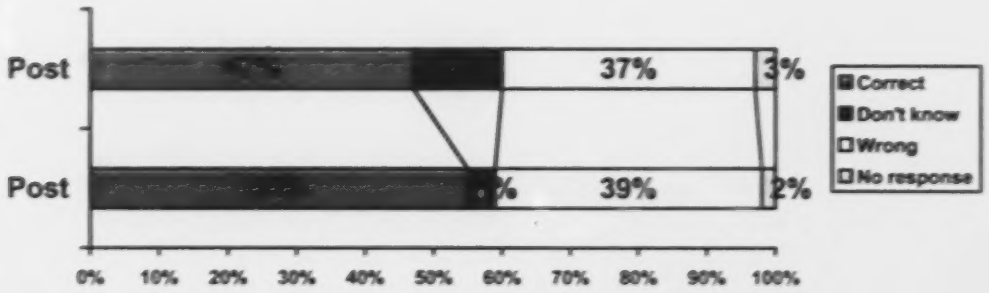
- For the pre-course data, perceived and actual knowledge were significantly and strongly correlated: $p = 0.000$, $r = 0.632$, $N = 39$.
- For the post-course data, the two variables are not significantly correlated ($p = 0.556$, $r = 0.097$, $N = 39$)
- For the learning data, perceived and actual learning were significantly correlated but less strongly than the pre-course data: $p = 0.001$, $r = 0.527$, $N = 39$.

[6] Self-assessed learning gains = difference between post-course and retrospective pre-course self-assessment. The ratings were converted to a scale of 100 using the formula: (rating-1)*25.
Actual learning gains = difference between post-course actual knowledge assessment scores and pre-course actual knowledge assessment scores. The scores were converted to a scale of 100 using the formula: (score/number of questions)*100.

Method III observed the participants' responses to the pre-post actual knowledge assessment from the point of view of their ability to perceive whether or not they knew the answer to a question. For this, we compared the participants who correctly assessed their knowledge of the question by selecting either the correct answer to the question or 'I don't know,' with the participants who selected a wrong answer to the question. Our data set included the responses of 160 participants in seven courses.

Even though the form given to the participants may look like a test, we explain to them that the process does not aim to test them. The questionnaires are anonymous and the results do not affect them personally. We explain to them that the process is an evaluation aiming to assess the effectiveness of the course. We encourage participants to be honest with their response and use the 'I don't know' option, when they think that they do not know the answer.

When participants answer a question, there are four possibilities in their responses: 1) choosing the correct answer; 2) choosing the 'I don't know' option; 3) choosing any of the wrong answers; or 4) skipping the question. Once the entire test is completed, any participant ends up with a combination of these four possibilities. Below is a chart indicating the average percentages of all 160 participants for each of these four categories to the pre-test and to the post-test.



We grouped together the correct responses with the 'I don't know' responses and treated them as one variable representing the proportion of the test where the participants correctly assessed their knowledge. The wrong answers represented the part of the test where they did not assess their knowledge correctly. The 'no responses' were treated as missing values. Based on these definitions, we wanted to know if some of the demographic groups available to us[7] were more likely than others to assess their knowledge correctly.

Since every participant was likely to have a combination of the four possible options on his/her test, we decided to arbitrarily divide the population into three groups:

a) those who were highly able to assess their knowledge correctly,
b) those who were poorly able to assess their knowledge correctly, and
c) those with an average ability.

BEST COPY AVAILABLE

[7] Identical to those described in Method I.

The groups were divided along one standard deviation above and below the mean. We observed cross-tabulations across the demographic groups with the different levels of participants' abilities to assess their knowledge correctly.

We found a significant relationship between the ability to assess knowledge correctly and work experience on the pre-test (N=156, p=0.001). However, this relationship was not linear. Among people with less than 3 years of experience (N=30) 80% had an average ability to assess their knowledge correctly, and 16.7% had high ability. Among people with 3 to 5 years of experience 82.4% had an average ability, and 8.8% had high ability (N=34). People with 6 to 10 years of work history (N=34) were more likely to have a high ability, 32.4%. For people with more than 10 years of experience (N=58), only 55.2% had average ability and 13.8% had high ability. As many as 31% of those with over 10 years of experience did not assess their knowledge very well.

We could not find this relationship replicated on the post-test. Besides this, we found no other relationship between the demographic variables observed and the ability to assess their own knowledge.

Conclusion

Contrarily to Nancy Dixon's findings, we found some significant relationships between what people felt they learned and what they actually learned. However, the relationships were weak and mostly explained by the pre-course data—where available.
We also found no consistent difference among the demographic groups that we observed: gender; educational attainment; region of origin; and the years of experience in the field of the course.
Therefore, we cannot rely on self-reported data to measure knowledge acquisition.

Limitations

This study is a work in progress with a relatively small data set at this point. We plan to keep looking at the relationship between perceived and actual learning as our data set increases, notably among sub-categories of our data set.

Before attempting any generalization, one should be aware of the specificity of our data. Our participants are coming from all over the world and are generally very well educated. We have very little variation between the less and the more educated among them, with very few of them having less than a Masters degree.

WBI training courses aim to help participants make better policy decisions. Skills' transfer is rarely the objective. Yet our Level 2 knowledge assessment method relies on multiple-choice questions, mostly because of resource implications. We recognize that alternative methods to measure actual learning should be developed.

Bibliography

Alliger GM, Tannenbaum S, Bennett W. JR., Traver H. and Shotland A. (1997) A Meta-analysis of the relations among training criteria, *Personnel Psychology*, 50, 341-358.

Dixon N. (1990) The relationship between trainee responses on participant reaction forms and posttest scores, *Human Resource Development Quarterly*, vol 1, no 2, 129-137.

Faerman, SR and Ban, C. (1992) Trainee satisfaction and training impact: Issues in training evaluation, *Public Productivity and Management Review, 16* 299-314.

Harrison, JK (1992) Individual and combined effects of behavior modeling and the cultural assimilator in cross-cultural management training, *Journal of Applied Psychology, 77*, 952-962.

Hilbert J., Preskill H. and Russ-Eft D. (1997) Evaluation Training, *What Works: Assessment, Development, and Measurement*, 109-150, the American Society for Training & Development.

Kirkpatrick D.L. (1994)Evaluating Training Programs – The Four Levels.

Mathieu, J.E., Tannenbaum, S.I., and Salas E. (1992) Influences of individual and situational characteristics on measures of training effectiveness, *Academy of Management Journal, 35* 828-847.

Warr P. and Bunce D. (1995) Trainee characteristics and the outcomes of open learning, *Personnel Psychology, 48*, 347-375.